MASTER'S THESIS

Incorporation of succesful Agile elements in Data Science Development processes
How a possible model is perceived by experts

Blasweiler, M.C. (Mariska)

Award date:
2019

Incorporation of successful Agile elements in Data Science Development processes

How a possible model is perceived by experts

Degree programme:	Open University of the Netherlands, Faculty of Management, Science & Technology Business Process Management & IT master's programme
Degree programme:	Open University of the Netherlands, Faculty of Management, Science & Technology Business Process Management & IT master's programme
Course:	IM0602 BPMIT Graduation Assignment Preparation IM9806 Business Process Management and IT Graduation Assignment
Student:	Mariska Blasweiler
Identification number:	
Date:	August 20th, 2019
Thesis supervisor	Prof. dr. ir. Remko Helms
Second reader	J. Baijens, Msc.
Version number:	1.0
Status:	final

Abstract

This research gives answer to the research question **How to incorporate the Agile methodology into Data Science projects to gain flexibility?** To get this answer a theoretical framework is built based on known Data Science development processes and Agile methods. The aim of the framework is to research which development process and Agile methods best fit together. This is a combination of CRISP-DM as the development process, KANBAN and SCRUM. The latter two are Agile methods. From this a SCRUMBAN model was created. This model was reviewed by consultants through a demonstration and interview. The results were then sorted and evaluated on four criteria: feasibility, completeness, usability and effectiveness. It was concluded that the model met all four criteria although effectiveness is only a perceived effectiveness. With little changes to the model, it is ready to use in practice to test if the model is effective.

Key terms

Data Science process, Knowledge Discovery, Agile, SCRUM, KANBAN

Summary

Data contains knowledge and knowledge is valuable. In the last decade companies embrace this value. Investments in Data Science Projects, Knowledge discovery and data have risen. Still a lot of these projects fail. The failure rate is 85 percent. The risks for companies are high when resources have been allocated or even used.

Reason for the failure of projects can be found in the process. It takes too long for projects to be researched, built and released. Traditional development processes focus on documentation and agreements. In recent years a new methodology was created to address this problem. The main focus of Agile is on delivering working software frequently. Reviewing the software with stakeholders gives feedback on a regular basis. An Agile development process is proven to be more flexible and adaptable to change. This means satisfied stakeholders and less risk for companies.

Data Science has several development models but two of them are the standard. Knowledge Discovery in Databases (KDD) which consists of sequential process steps. And CRoss Industry Standard Process for Data Mining (CRISP-DM), which has an iterative nature. This means that previous executed process steps can be executed again when needed. Through the years several attempts were made to incorporate Agile models to the existing Data Science Development models. Still the problem of high failure rate for Data Science Projects exist.

To construct a theoretical framework a systematic literature research is conducted. From the possible Data Science development models, the CRISP-DM process model is selected. Next step is to compare suitable Agile methods. The research of these methods is based on previous attempt fit Agile methods to Data Science projects. The literature shows attempts with SCRUM and KANBAN. SCRUM provides a complete framework with roles, events and artefacts. To make it suitable for Data Science projects, the flexibility of KANBAN was added. KANBAN creates a continuous workflow of work-in-progress.

For the research the Design Science Research Method is used. The choice for Design Science Research Method is due to the practical nature of this research. This method provides a step by step framework for performing a rigorous and relevant research. First the problem and motivation are defined. Secondly, objects for the artefact are defined from the theoretical framework. And finally, the artefact is designed and created. In this case a model that incorporates Agile elements into the CRISP-DM model: the SCRUMBAN Data Science development model. Scientific rigor is guarded through recordings of the taken steps, decisions and results.

The model is created based on SCRUM. Roles, events and artefacts are defined. The complete process is defined by using these roles, events and artefacts. From KANBAN the board and the workflow are added to the model. The model was then finished and ready for evaluation. The evaluation is done on four evaluation criteria: feasibility, completeness, usability and effectiveness. The evaluation is conducted at a large consulting company based in the Netherlands. The consultants were selected based on Data Science and Agile experience. Another selection criterion is the role they currently execute. The interviews were held individually and recorded for evidence. The recordings were translated into scripts and form the base for the results. These results were divided into either positive remarks that support the model or negative remarks and points to improve. The positive remarks were made on the feasibility and effectiveness of the model. The

consultants didn’t see any impediments for the model to be tested in a next phase of the research. They thought it was practical and clear. On completeness there were multiple remarks made to adjust the roles assigned to a team. The initial version of the SCRUMBAN Data Science development model contained an architect within the development team. After discussion it was concluded this was not feasible, but it does add value. Recommendation was to make the architect a consulting role. Another recommendation came through the evaluation of usability. The consultants all questioned the points system from the initial version of the SCRUMBAN Data Science development model. The conclusion was drawn, that the point system is to complex and should be replaced with a simple weighing method. The suggestion to use a wall-of-reference was added to the SCRUMBAN Data Science development model. This is the second version of the model. The conclusion to the research question is, that the model contains elements of Agile that fit Data Science projects. This helps the team to gain insight in work-in-progress and possible changes. Therefore, the SCUMBAN Data Science development model provides a framework with Agile elements and the model adds flexibility to Data Science projects.

Contents

Contents

1. Introduction

1.1. Background

Data science is the discovery of knowledge from data and the presentation of this knowledge. There are two sides to the discovery of knowledge. The first is that companies have a specific problem they want a solution for. The second is the hidden gems in a dataset that are discovered through Data Science. This means that the business doesn't have a clear requirement or goal. This discovery of knowledge could contain an advantage towards competitors in the market (Dhar, 2013). Therefore, Data Analytics, Embedded Analytics, Business Intelligence and Data Mining are subjects of Data Science for professionals to read about. The fear of missing out on this advantage causes companies to invest in data, data science and knowledge from data. Data Science has been around for decades but the valuation of Data Science as a strategic asset is relatively new. The first literature dates back to the 60's with the first publication about 'Datalogy', a new term for science on data and data process (Naur, 1966). Though it seems the subject is about data science, the focus is on computer science and that doesn't change till the 90's. In the 80's and 90's data has a more statistical approach in the form of Business Intelligence. Reporting and data of past events give companies insights to make strategic decisions. Companies gained insights from data and started to collect data to achieve an advantage over their competitors. Through the years data has evolved to possess predictive capabilities and insights. These capabilities and insights contain value. To increase that value and have an advantage over competitors, companies are collecting massive amounts of data, building data lakes and are starting to develop their own data science projects (Cao, 2015). There are several methodologies and processes for Data Science projects. For example, the classical waterfall process model and Agile development models.

The Waterfall process model is a classical development model and is also known as the Systems Development Life Cycle (Royce, 1987). As the name implies the process model is used for System Engineering, nevertheless it is also applied to Data Science projects. The model consists of several phases that need to be completed before going on to the next phase in the process. A step is completed when all stakeholders agree on the outcome. For example, a blueprint of the proposed solution is accepted by the business. The process starts with defining the requirements of the business, followed by a detailed architectural and development design. The emphasis is on documentation of the requirements and the designs. These must be approved before development starts. There are two risks involved with the Waterfall process model. The first risk is that the delivered product does not meet the expectations of the stakeholders. The second risk is that the delivered product is invalid by the time it is finished and released. These risks emerge when the requirements change over time or there are other changes that affect the product (Avison & Fitzgerald, 2003).

In 2001 the Agile manifesto was written to address the risks of the Waterfall model. The emphasis of Agile is on people, customers, working products and flexibility. Agile development models are iterative which means that steps can be repeated if needed. Working products are delivered and reviewed frequently. The review sessions give the stakeholders the possibility to give input and the development team to adapt to changes quickly resulting in satisfied customers (Beck et al., 2001).

Software development projects adopted the new way of working as the holy grail to improve on time-to-market. The Agile way of project execution has been researched on effectiveness. Evidence was found that Agile models possess a positive effect on project success (Serrador & Pinto, 2015).

A similar shift from Waterfall to Agile can be detected in process models for Data Science projects. It started with Knowledge Discovery in Databases (KDD) as the initial process model and this evolved to CRISP-DM, a more iterative development process model (do Nascimento & de Oliveira, 2012; Mariscal, Marbán, & Fernández, 2010). The reason for this shift can be found in the failure rate of Data Science Projects where 85 percent of all Data Science Projects fail. The cause of this failure rate is that not all expectations are met, they don't reach the set goals, requirements or objectives. This is due to the problem that projects aren't flexible to adapt to changes in requirements (Asay, 2017; do Nascimento & de Oliveira, 2012; Walker, 2017). The same reason was given for the failure of the waterfall method. Agile methodologies are proven to be successful for software engineering. Agile development processes prove to be more flexible and can adapt to change quickly. Now attempts for a more Agile approach for Data Science development processes are introduced. Still companies struggle how to incorporate Agile into the Data Science development process.

1.2. Problem statement

As described in the previous paragraph Data Science Projects have a high failure rate. There is a need for a development process that gains insights on a frequent basis and quickly adapts changes.

1.3. Research objective and questions

Based on the stated problem a more Agile approach for the development process could be the solution to gain more flexibility. Agile development models are proven solutions for System Engineering. They support developers to be more flexible, gain insights quickly and adapt changes to the working product. Agile development models could have the same advantage for Data Science projects (Serrador & Pinto, 2015). The objective of the research is to develop a Data Science process model based on Agile and Data Science development models. The research question that arises is: **How to incorporate the Agile methodology into Data Science projects to gain flexibility?**

To answer the question several sub questions were developed:

SQ1. What are Agile values and methods and which one(s) will fit Data Science development processes? This question will be answered by literature study.

SQ2. Which Data Science development models are available, which attempts have been made to incorporate Agile into the development models? This question will be answered by literature study and will result in a proposed Agile Data Science development model.

SQ3. Is the proposed model suitable for Data Science projects? This question will be answered by Design Science Research containing a model development, demonstration and interviews with experienced consultants.

1.4. Motivation/relevance

Research of several Data Science processes show the evolution of models but stops before Agile was introduced to the process models (Mariscal et al., 2010). In Software Engineering Agile has proven advantages over classical waterfall (Saltz & Heckman, 2018). Attempts have been made to incorporate Agile methodologies to standard Data Science processes (do Nascimento & de Oliveira, 2012; Larson & Chang, 2016) but a complete model with organization of the process is not available yet. This research will contribute to the existing research by development and testing of a complete model that describes how to incorporate Agile into existing Data Science process models.
This research will result in advice on how businesses can use Agile methods or parts of it to improve their Data Science development process. Businesses will gain insights on feasibility of these projects in an early stage and adapt to it, this will improve flexibility and customer satisfaction for these businesses.

1.5. Main lines of approach

The research is mainly exploratory (Saunders, Lewis, & Thornhill, 2016). At first a literature study is done on Data Science development processes and the Agile methodology. The method used for the literature study is a systematic process. The literature study will give answer to two sub questions: which Agile values and methods fit the Data Science development process and what attempts have been made to incorporate Agile methods into the Data Science Development process. This constructs the theoretical framework. From this theoretical framework elements of Agile development methods are selected to be incorporated into the Data Science process model. To develop and test the model the Design Science Research method is used. In the testing phase the model is presented to stakeholders from the field of Data Science. The feedback gives the data that forms the base of the conclusion. In the end suggestions for further research is presented.

2. Theoretical framework

This section provides the theoretical framework. First the research approach will be described, this will give structure how the research will be executed. The second paragraph shows the execution of the research approach. And in the third paragraph follow up research is discussed.

2.1. Research approach

The aim of the literature study is to give answer to the question:
What are Agile values and methods and which one(s) will fit Data Science development processes?

The systematic process for literature research is based on the eight step guide (Okoli & Schabram, 2010):
Step 1: Purpose of the literature review
Step 2: Protocol and training
Step 3: Searching for literature
Step 4: Practical screen
Step 5: Quality appraisal
Step 6: Data extraction
Step 7: Synthesis of studies
Step 8: Writing the review.

2.2. Implementation

2.2.1. Purpose of the literature review

The purpose of the literature review has been written in the introduction of the paper. It is used to create a theoretical framework by answering the research questions in previous chapter. The theoretical framework supports the proposed model.

2.2.2. Protocol and training

The second step is protocol and training, it is one of the planning steps together with the purpose of the literature review. The protocol is a plan how to execute the literature review and consists creating the research question. The steps of the protocol are written in the previous chapter. Training is not executed because there is only one researcher and alignment between different researchers isn't required.

2.2.3. Searching for literature

When the planning is complete the third step is executed. Searching for literature is started by the development of keywords. The keywords are linked to the research questions for the previous chapter. The first keywords are the nouns from the research questions. Next is to add substitutions of the first keywords to extend the query. The keywords used for literature research are: Data Science, Knowledge Discovery, Data Mining, Data Science models, development process, development models, Iterative process. For combinations of these keywords the command AND is used and for substitutions of keywords the command OR is used to enter the queries searched on.

In the table below the keywords and substitutions are displayed.

Keyword and substitutions (AND)	OR	Keyword and substitutions (AND)	OR	Keyword and substitutions (AND)
Agile		"Data Science"		Process
Iterative		"Knowledge Discovery"		Method*
		"Data mining"		Model
		"Big data"		Framework
		"Data Analytics"		
		"Business Intelligence"		

The keyword Agile is used for search on all the words in an article, this is done to assure that all articles with Agile will be considered for the results even though it is not mentioned in the abstract. The other keywords and substitutions are searched in the abstract. If the words are not in the abstract, they are considered not relevant for the article.

The source for the scientific literature is the library of the Open University Netherlands. The settings for the query are that the results are articles which were peer-reviewed and written in the English language. The results are sorted on relevance. When the results are filtered step four to six will be conducted. The execution of the query resulted in 739 hits.

2.2.4. Practical screen

The practical screen is used to exclude the papers that are not relevant for the theoretical study. In this step the results are checked on the following criteria:

1. Restrict to the first hundred, since the sorting is on relevance the first hundred articles are considered to be the most relevant.
2. If the keywords in the titles gave suspicion to answer the research question.
3. Whether the content of the article contributes to the theoretical framework.

The first one hundred results were selected for review on the title, as they were considered most relevant based on the sorting of the results of the query. If the title was expected to answer the questions, the paper was selected for the next step. The eighteen articles that are selected, are considered most relevant.

2.2.5. Quality appraisal

The articles considered for the theoretical framework are qualitative studies. They contain research done on methods and models for Agile and Data Science. To rate the articles on quality, the articles were researched on argumentation of the conclusions. The articles that were supported with theory or a case study were of better quality than the conclusions based on supposition.

2.2.6. Data extraction

Data extraction will be executed by retrieving the information from the articles that are included in the previous step. To extract the correct data for the theoretical framework a connection to the research question is made. The data extracted from the articles contains information about Data Science models, Agile methods and attempts on Agile for Data.

From the eighteen articles considered for the theoretical framework selected in the practical screen, seven were regarded to be of good quality for this research. For an overview of the selected articles

see appendix I. The extracted data is theory, conclusions, limitations and suggestions for further research.

2.2.7. Synthesis of the studies

After the data has been extracted from the articles, it is divided into subjects. The connections between data and sorting, belongs to the step synthesis of the studies. The first subject is the data science process. This is data that explains about the different processes that exist for Data Science projects. Another subject is Agile and Agile methods, this describes what Agile is and which methods exist. The last subject is on Agile and Data Science. Data about attempts of Agile methods used in Data Science projects and results are connected to each other.

2.2.8. Writing the review

The final step of the Structured Literature Review is writing the review. The review of the articles and results are presented in the next paragraph.

2.3. Results and conclusions

2.3.1. Data Science process

The data science process originated from the Knowledge Discovery in Databases KDD (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). This process consists of the steps: data preparation, data mining, discovery of patterns and interpret/evaluate knowledge. From this process several other models and methodologies were created such as SEMMA and Two Crows. When the other models and methodologies are compared to KDD they show similar process steps to be taken although they have different names (Mariscal et al., 2010). The issues with these models and methodologies is that they are bound to an industry, tool or application. To solve this issue a new methodology: CRoss Industry Standard Process for Data Mining CRISP-DM (Chapman et al., 2000) was created. This methodology is considered the standard process for Data Science projects. This process consists of the steps: Business understanding, Data understanding, Data preparation, Modelling, Evaluation and Deployment.

Different to KDD, CRISP-DM has the steps Business understanding and Data understanding. The first step is Business Understanding. In this step the focus is on project objectives and requirements. And the outcome of the step is a plan how to address the objectives and requirements with a problem definition. The second step to get familiar with the data and to get the first insights in the data or hidden information. The process has steps that are iterative, they can be repeated when needed. When in the data understanding phase, the goal can't be achieved, it is possible to execute the first step again. When data understanding is completed the process proceeds to the data preparation phase. A final data set is created in this step. In the modelling phase several modelling techniques can be applied. It is possible to go back to data preparation for some techniques require specific data sets. When the model is created it is evaluated. Two outcomes are possible at this phase in the process: there are new requirements on the model and the process starts again at the first step or the model is deployed for use, which ends the development process.

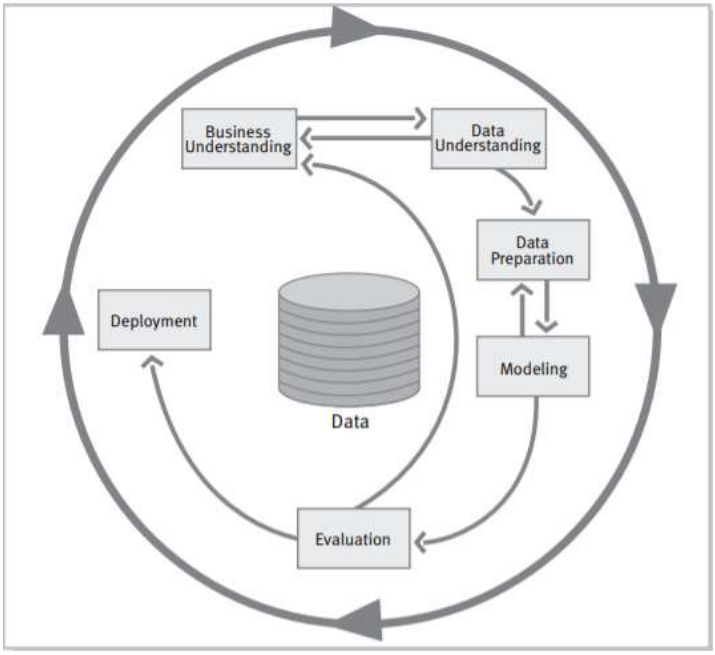


Figure 1: Phases of the CRISP-DM reference model (Chapman et al., 2000)

Although CRISP-DM is considered to be the standard, it seems it has not reached the maturity to address complex Data Science development requirements such as project management of multidisciplinary teams (Mariscal et al., 2010). A solution for the immaturity of CRISP-DM is to incorporate organizational activities such as project management from mature Software Engineering SE methodologies and processes. Although the attempt of combining CRISP-DM with mature SE processes it is mentioned that process life cycles were not considered. The life cycles mentioned are waterfall, incremental and iterative life cycles (Marbán, Segovia, Menasalvas, & Fernández-Baizán, 2009). Though waterfall is mentioned as a mature life cycle several studies show that waterfall projects are not effective (Avison & Fitzgerald, 2003; Kisielnicki & Misiak, 2017; Serrador & Pinto, 2015).

2.3.2. Agile

To address the issues of waterfall projects the Agile manifesto was written. The Agile manifesto consists of four values and twelve principles. These values and principles support the shift from classical waterfall development which is process based to an iterative development which has the focus on the customer and de development team. The four values are (Beck et al., 2001):

1. Individuals and interactions over processes and tools
2. Working software over comprehensive documentation
3. Customer collaboration over contract negotiation
4. Responding to change over following a plan

These values are supported by twelve principles:

- Satisfy the customer through early and continues delivery of valuable software
- Welcome changing requirements, even late in development. Agile processes harness change for the customer's competitive advantage.

- Deliver working software frequently, from a couple of weeks to a couple of months, with a preference to the shorter timescale.
- Businesspeople and developers must work together daily throughout the project.
- Build projects around motivated individuals. Give them the environment and support they need and trust them to get the job done.
- The most efficient and effective method of conveying information to and within a development team is face-to-face conversation.
- Working software is the primary measure of progress.
- Agile processes promote sustainable development. The sponsors, developers, and users should be able to maintain a constant pace indefinitely.
- Continuous attention to technical excellence and good design enhances agility.
- Simplicity--the art of maximizing the amount of work not done--is essential.
- The best architectures, requirements, and designs emerge from self-organizing teams.
- At regular intervals, the team reflects on how to become more effective, then tunes and adjusts its behaviour accordingly.

The researchers concluded that Agile will improve collaboration between the development teams and the business. This collaboration will improve quality of the requirements and therefore, improve the quality of the delivered product (Larson & Chang, 2016). The reason why Agile is a better fit than waterfall is due to the issue that people do not know what they want. However, they can indicate on the basis of a presentation what their aim is (Kisielnicki & Misiak, 2017). In classical waterfall the aimed product is set in stone and the build phase starts. In the end the results are presented to the end-user while they were not involved during build. When changes are indicated in the end, these are expensive and time consuming. It is concluded that Agile shows value in the Data Science process because required changes become clear at an earlier stage and stakeholders are involved from the beginning (Kisielnicki & Misiak, 2017).

The main Agile development methods are: Crystal methodologies, Dynamic Software Development Method, Feature-Driven Development, Lean Software Development, SCRUM, KANBAN and Extreme Programming (Dybå & Dingsøyr, 2008). These methods all focus on fast and frequent delivery of solutions, human interactions and reduction of steps that don't add value to the product. From these methods SCRUM, KANBAN and Extreme Programming are considered to be the most popular used Agile methodologies used in Software Development, Business Intelligence and Data Science (Larson & Chang, 2016; Muntean & Surcel, 2013).

2.3.2.1. SCRUM

SCRUM is a methodology that consists of a team, events and artefacts (Schwaber & Sutherland, 2017). The team, events and artefacts are building blocks of the complete process and are not optional. All team members should have sufficient knowledge of the methodology before starting a development.
The roles in a SCRUM team are:

The product owner: the representative of the business. This person is responsible for optimizing the value of work of the team and the backlog.

Development team: the team that creates working products. A team should be small enough to act quickly and large enough to get work done. This results in teams between three and nine members. An important aspect of the development team is that they are self-organizing, cross-functional and without hierarchy.
SCRUM master: The SCRUM master is a supportive role for the product owner, the development team and the business. This person helps everyone to understand Scrum and helps to maximize value of the team.

The events of Scrum are set and used to create regularity. All events have a pre-set time-box, which means that every event has a set duration. It is emphasized that all the events should be included otherwise this will result in a decrease in effectiveness of the methodology.
Sprint: The complete loop in which a product increment is created. In the sprint all team-members, artefacts and events are included. The complete sprint is time-boxed with a maximum of a month. A sprint may only be cancelled by the product owner since this person is the representative of the business. During the sprint no changes to the scope are allowed unless re-negotiated with the product owner.
Sprint planning: The work to be done in a sprint. The list of work to be done is created by the entire Scrum team. Time-box for the planning is eight hours maximum.
Sprint goal: The objective of the sprint and helps the team to gain insight for the importance of the increment. It is part of the Sprint planning.
Daily Scrum: A fifteen-minute meeting of the development team to gain insight of the progress of the sprint.
Sprint review: The development team shows the created increments of the sprint. The delivered increments and the backlog form the base of the new sprint planning. The review is time-boxed for a maximum of four hours.
Sprint retrospective: The Scrum team inspects itself and creates improvements on the team to be implemented the following sprint. This event is time-boxed for a maximum of three hours.

Scrum artefacts are:
Product backlog: The complete list of everything to be needed in a product. It is the only source of requirements.
Sprint backlog: A list of items selected to be developed in the sprint.
Increment: The sum of all Product Backlog items completed during a sprint.

Figure 2 shows the complete process with all the SCRUM elements.

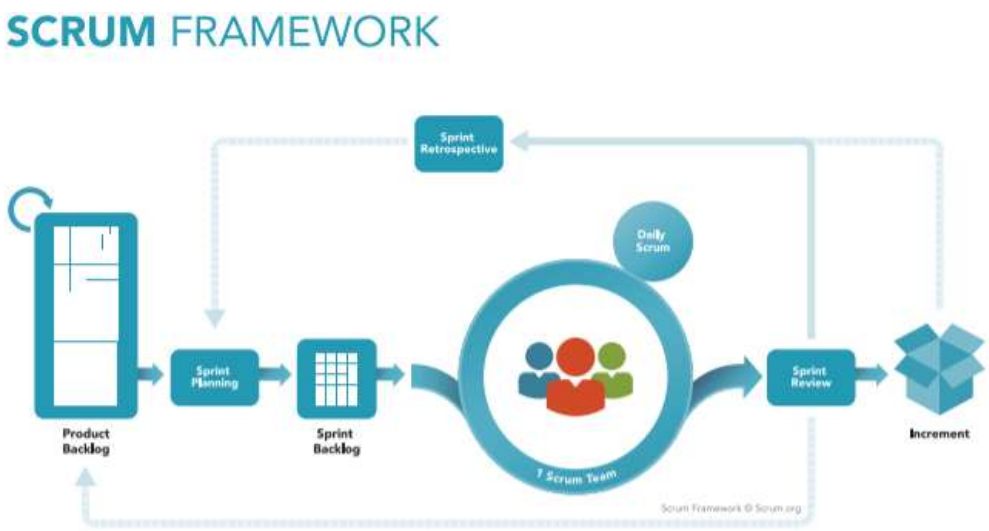


Figure 2: SCRUM framework (scrum.org)

For Software Development SCUM is a proven methodology, with set goals that can be achieved.

2.3.2.2. KANBAN

KANBAN is based on LEAN software development. The main focus of LEAN is to cut ineffective activities from processes and keep value added activities. KANBAN consists of four values:

- Visualise workflow
- Limit Work-In-Progress
- Measure and manage flow
- Make process policies explicit

Through a KANBAN board the four values are transparent for the team. On this board Work-In-Progress is show with tasks and steps. Whenever a step is completed the task is moved to the next step until it results in done. A maximum of tasks is set according to the size of the team. The team is self-organized and cross-functional. The focus of KANBAN is continues delivery and team empowerment (Saltz & Heckman, 2018)

Figure 3: Kanban process overview (Lei, Ganjeizadeh, Jayachandran, & Ozcan, 2017)

2.3.2.3. Extreme Programming

The aim of Extreme Programming is customer satisfaction (Wells, 2013). This is done through adaptability of the projects when requirements change. It is important that changes can be made even late in the development life cycle. Furthermore, Extreme Programming emphasizes on

teamwork, everyone in the process become equal partners to solve problems efficiently. The teams organize themselves and don't need to be appointed by management.

Five principles of Extreme Programming:

- Communication
- Simplicity
- Feedback
- Respect
- Courage

The steps of Extreme Programming are kept simple in a flowchart. Activities that don't contribute to the process are excluded to reduce inefficient use of available recourses. Figure 3 shows the flowchart of the iteration step in the complete flowchart of Extreme Programming.

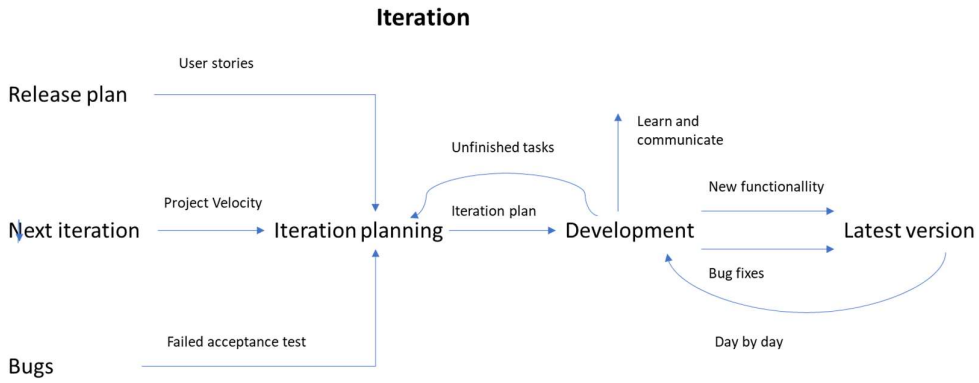


Figure 4 Iteration in Extreme programming (Wells, 2013)

2.3.2.4. Differences and similarities

Although SCRUM, KANBAN and Extreme programming are all Agile processes, they are not completely the same. Where SCRUM and KANBAN visualize the process through boards, Extreme Programming focus is on simplicity in a workflow. Different to SCRUM and Extreme Programming KANBAN doesn't have a clear process. The steps or phases are created by the teams and work-in-progress is pulled through the steps. The focus of KANBAN is to create a continuous flow of work-in-progress that can be handled by the development team. In comparison to KANBAN and Extreme Programming, SCRUM has set rules for time-boxed events, team roles and artefacts. Team members should know the rules of SCRUM and it takes time for a team to gain speed.

On the other hand, all the methods have the same goal: Customer Satisfaction. In all of them the customer is part of the team. Communication is key and regular meetings to know what the status is of a development. Another similarity is self-organizing teams. This helps to solve issues quickly.

2.3.3. Agile Data Science process

In the beginning the Agile values and principles were mainly focussed on software development. When it became clear that there are similar issues found in Data Science projects as in software

projects, Agile was considered to be the solution. One of the researches focused on the supporting tools for Data Science projects in an Agile environment. It was concluded that the supporting tools for data mining and discovery were not suitable for Agile Data Science Projects (Grigoriev & Yevtushenko, 2003). One of the conclusions was that the process is finished when a model is built, which makes it impossible to add changes to the model. To adapt changes a new model must be created. This makes an iterative process impossible to execute. Another conclusion is that the human aspect of Agile isn't addressed because modifications to algorithms isn't possible in any of the supporting development tools. Thus, it becomes clear that at least three of the four values of the Agile manifesto can't be met. There isn't any conclusion about the contract value, and it seems irrelevant for the research on supporting tools. An important conclusion from the same research is that they didn't expect the supporting tools to be suitable in the near future. Evidence for this conclusion can be found in research done on supporting tools for Agile Business Intelligence (Muntean & Surcel, 2013). Business Intelligence is not the same as Data Science although it shows some similarities. Business Intelligence is the ability to create knowledge from data. These are reports that give insight on past events. Data Science takes this data and creates predictive information. Data Science requires a different approach and is more subject to change than Business Intelligence. (Larson & Chang, 2016). To be able to adapt to change quickly it is important to have an Agile development process and an Agile support system as well (Muntean & Surcel, 2013). Though the tools do not support the Agile development methods for different aspects of Data Science, it is marked by different researchers that the Agile development methodologies are a better fit for Data Science processes than the classical waterfall development method (Kisielnicki & Misiak, 2016; Larson & Chang, 2016; Muntean & Surcel, 2013).

In order to use the value of Agile for the Data Science process attempts have been made to incorporate Agile life cycles, principles and methodologies into the processes KDD and CRISP-DM. For example AgileKDD was researched (do Nascimento & de Oliveira, 2012). Agile KDD is a framework for a KDD and BI development processes. The life cycle of the process consists of four steps: Inception, Elaboration, Construction and Transition. Each step has sub steps that need to be taken. These sub steps are derived from KDD and CRISP-DM. Agile elements were added to the process such as small work units, incremental development through small cycles, continues feedback due to the small cycles and demonstration to stakeholders. Eventually a working application is delivered. The process was tested with a case study. Adjustments and improvements on the model were needed but details about the adjustments and improvements were not mentioned.

In later researches visualization was added to the process resulting in the Agile delivery framework (Larson & Chang, 2016). The proposed process consisted of the steps: Scope, Data Acquisition/Discover, Analyse/Visualize, Model/Design/Develop, Validate and Deploy. Due to the iterative nature the Data Science process is perceived to be Agile. Iterations are mostly found in the first four steps. From the validation step changes could emerge but the step itself is not iterative. Like AgileKDD this process is executed by small teams with collaboration between business and expert developers. The difference remarked between both processes is that AgileKDD suggest time-boxed increments while for the Agile delivery framework, this time-box is not needed. The main objective of the process is to create an analytical model with the best results.

One attempt for adding Agile practices to Data Science Development processes has been researched. Agile practices continuous integration, test-driven development, pair programming, user stories and SCRUM events were found to align easily into the KDD process which was extended with the Business understanding and Data understanding steps from CRISP-DM. The effect on understanding what the user needs and a better understanding of the objective was noticed and

therefor delivering products with value for the business (Schmidt & Sun, 2018). Though this research is concluded to have positive advantages it doesn't show which Agile practices give the most benefits. Furthermore, the question rises how the model will fit cross-functional teams. Applying parts of scrum is in contrast with the statement that the complete methodology needs to be implemented otherwise transparency will be affected and result in a less effective methodology than it could be.

In a case study conducted with students executing data science projects with Agile or iterative based methodologies the performance of these processes was measured. The performance was measured on reaching goals (effectiveness) and adoptability of the process (efficiency). In this case study one Data Science Development process (CRISP-DM) two Agile methodologies (SCRUM and KANBAN) are applied. The case study concluded that CRISP-DM and KANBAN performed well for the groups and they outperformed SCRUM based projects. KANBAN didn't need as much explanation as SCRUM for a team to start working. CRISP-DM performed well on the systematic process and the clear steps to get through the development process. The group preferred quick adaptability of the KANBAN method and making the Work-In-Progress visible. The methods were measured how they performed on managing project schedules (Saltz & Heckman, 2018).

2.3.4. Conclusion

The table below gives an overview of the results of the literature review per section. This forms the basis for the conclusion described in this paragraph.

Section	**Results**
Data Science Process	The two standard processes for Data Science Projects are CRISP-DM and KDD. CRISP-DM has evolved from KDD and has an iterative nature. This means that developers can go back to previous steps when needed. The iterative nature gives the possibility to adapt Agile elements to the process. The KDD process consists of steps that are more like waterfall. Agile has the focus on communication, involving the business and show results on a frequent basis. KDD doesn't have the same goals. On the other hand, CRISP-DM focusses on Business Understanding and Data Understanding. There is more focus on the business and delivering on customer satisfaction.
Agile	The three processes SCRUM, KANBAN and Extreme Programming have similarities and differences. The similarities are proven to be effective. The focus should be on teams, frequent events and communication, business as part of the team and short delivery cycles. The difference can be found in the efficiency and effectiveness of the processes. KANBAN gives the team insights on progress and it helps not to take on too much work that slows the team down. In SCRUM there is a complete framework with events, teams and artefacts. There are insights in work to be done and commitment from the whole team. The team can dedicate themselves to the work to be done without being delayed by distracting activities. Extreme Programming is relatively easy to understand with focus on communication and customer satisfaction. A negative side to

	Extreme Programming is the lack of insights in the status of the work.
Agile Data Science Process	The results consist of three insights: The first is that not only the process should support an iterative way of working but the supporting software should be capable to adapt to change quickly. In the first attempts the software couldn't meet this requirement. The second insight is an attempt for an Agile KDD process. Parts of SCRUM are added to the KDD process and were concluded to have a positive effect on the process. Business understanding improved and that had a positive effect on delivering products that add value for the business. However, what elements of SCRUM were added is unclear. Another point is that other Agile methods are not combined with KDD or CRISP-DM such as KANBAN or Extreme Programming even though both are interesting to incorporate and test on effectiveness. And last but not least, was a case study on efficiency and effectiveness of SCRUM, KANBAN and CRISP-DM. KANBAN and CRISP-DM are found to be easy to understand and apply, but SCRUM requires better understanding of the process, teams, events and artefacts. Nevertheless, it should be mentioned that the cycle could be repeated more often to get real understanding of the efficiency and effectiveness of all methods.

Agile values and methods can fit Data Science development processes. The standard process CRISP-DM is an iterative process: not all steps should be completed to go to next step and going back to previous steps are supported. The iterative nature is preferred to align with the iterative nature of Agile, as this adds flexibility to the process. The Agile elements mentioned by researches is working in small teams, collaboration between business and expert developers. Though time-boxed steps like sprints in SCRUM could be incorporated not all researches agree to do this. Nevertheless, some sort of development management is needed, and visualisation of Work-In-Progress supports the projects. From these findings a model combining CRISP-DM with SCRUM and KANBAN elements could be developed for further research. The CRISP-DM model provides the process for Data Science projects, this process is standard and well known. It is iterative which fits Agile methodologies. From SCRUM organization elements are added to involve business with the development team. Demonstrations of working software to the business are part of the model. From KANBAN the board with the process steps of the development is used. Time-boxed projects are not viewed valuable for Data Science projects due to the uncertain outcome of the first four steps of the process. SCRUM requires to have a review of working software at the end of the time-boxed sprint. Therefore, an overview of Work-In-Progress gives insights quickly and reviews of working products is one of the process steps that is covered in CRISP-DM in the evaluation step.

2.4. Objective of the follow-up research

The objective is to create a model based on the mentioned CRISP-DM process with SCRUM and KANBAN elements. Since SCRUM and KANBAN are Agile methodologies, the model will show how Agile can support Data Science processes, which steps to take, how the process model is organized and the process model life cycle. The case study will test performance of the model and deliver recommendations for adjustments and improvements of the model.

3. Methodology

This chapter describes which research methods are used to build through a conceptual design. This will elaborate what and why it's done. The second paragraph describes how the research is executed. The next paragraph will indicate how the data analysis is executed. The chapter will end with a paragraph that reflects on the designed research approach.

3.1. Conceptual design: select the research method(s)

How Agile can be incorporated in Data Science projects, is answered through a proposed model. The model will be tested through empirical research. The research method used is Design Science Research Method (DSRM). The method is chosen because it helps to give answer to a design question instead of a knowledge question. The model can be improved by iterations. The iterations are executed through evaluation done by a target audience of experts in the field. DSRM provides a framework for execution of the research and consists of six activities:

1. Problem definition and motivation
2. Define the objectives for a solution
3. Design and development
4. Demonstration
5. Evaluation
6. Communication.

DSRM involves as stated by Peffers, Tuunanen, Rothenberger, & Chatterjee (2007) "A rigorous process to design artefacts to solve observed problems, to make research contributions, to evaluate the designs, and to communicate the results to appropriate audiences". An artefact could be a construct, model and instantiations. For this research a model will be created based on the theoretical framework. This model is going to be tested through interviews with experts. These are experts from the field of Data Science Development such as Managers, Business representatives, developers and Architects. Interviews are held due to available resources such as time. The results from one cycle will not provide enough data to answer the research question. Several cycles should be executed to give objective data. In order to test the model, processes should be adjusted. Available resources for execution of the research is not sufficient to change process and test more than one cycle. Interviews based on demonstration of the model will give data quickly and efficiently.

3.2. Technical design: elaboration of the method

In this section the technical design according to the activities from DSRM is described.

3.2.1. Problem definition

The problem definition and motivation are stated in the first chapter of this research paper. The problem is that Data Science Projects have a fail rate of 85%. Companies take risks by starting projects and allocating resources. Agile models are proven to be effective solution for similar problems in System Engineering. The same results could apply to Data Science processes.

3.2.2. Define the objectives for a solution

The objective of the solution is to create a model that incorporates Agile into Data Science projects. The theoretical framework provides data of previous attempts and conclusions on Agile Data Science processes. The model consists of the CRISP-DM model combined with SCRUM and KANBAN elements. The model is called the SCRUMBAN Data Science development model and contains roles, events and artefacts.

3.2.3. Design and development

To design the SCRUMBAN Data Science development model, successful elements of SCRUM and KANBAN from the theoretical framework are selected. These elements are incorporated into the model. The model provides a framework with a Data Science Development process and successful Agile elements. The model should be complete to demonstrate in the next step.

3.2.4. Demonstration

In this step the proposed model is demonstrated to five people with different roles in different domains. The demonstration is done in a personal meeting or via a digital meeting. Preference is to execute the demonstration in a personal meeting in which also unspoken information can be collected such as body language. When this is not possible the demonstration is done via a digital meeting. All interviews are recorded. This gives the ability to focus on the demonstration. Afterwards the data is extracted from the recordings.

All roles interviewed are selected on their experience with Agile projects, they should have at least participated in one Agile development. To get different perspectives the roles interviewed should differ to give more insights. The project manager role from the Software Engineering domain is added to give input on the model with deep Agile experience with at least five projects in an Agile setting.

The demonstration is recorded with a voice recorder. This has two advantages: first the demonstration is not interrupted while writing down the remarks. Second the recording is marked as evidence for the research.

For the demonstration a presentation will be built to guide the people through the SCRUMBAN Data Science development model. The presentation starts with background of the research and the research questions. Then it addresses Agile with the values and principals. After this the elements of the SCRUMBAN Data Science development model is presented.

3.2.5. Evaluation

After the demonstration the model is ready to be evaluated. The evaluation is conducted through semi structured interview questions. The interview is designed to keep track of the elements to reflect on. The interview questions for the different domains can be viewed in Appendix II. This will create some flexibility to adjust to circumstances when needed (Saunders et al., 2016). The collected data from the demonstration is the input for the evaluation of the proposed model. The collected data is unstructured. The individual remarks are evaluated as positive or negative. The positive remarks support the advantages of the created and evaluated model. The negative remarks will result in adjustments and improvements of the model.

3.2.6. Communication

The executed research and results will be presented. Publication of the research is done by the university when the graduation report is considered sufficient. Other communication is not relevant for this research.

3.3. Reflection w.r.t. rigor and relevance

Scientific rigor is the ability to replicate executed research. Design Science Research Method provides a framework for the execution of the research. Choices and outcomes of every step of the research is documented. An artefact is created on basis of the theoretical framework. The first step of the framework is problem definition and motivation. The aim of Design Science Research is to contribute to existing theory by creating and testing an artefact. In this research the artefact is a model. The model will address the problem stated in the first chapter. The relevance of the design research method is creating a model to reduce the failure rate of Data Science Projects.

The problem definition and motivation therefor provide relevance of the aimed research. In the second step objectives for a solution are defined. The objectives are found through systematic literature research. The steps, choices and results are documented in chapter two. The artefact is based on the results of the theoretical framework. The artefact and research therefor contribute to existing theory.

The aimed artefact for this research is a model based on Data Science Development Processes and Agile elements. The elements are found in the theoretical framework. The relevance of the research is considered when creating the model. How the model is created, and which choices are made is documented for replication of the research.

The experts are selected from existing stakeholders in the field of Data Science. How the selection of stakeholders is aimed, and the execution shall be documented. For privacy reasons only functions will be documented. The model is demonstrated to these experts. Together with the demonstration an interview (evaluation) will be held. The interview is semi-structured. To excluded steering in the answers from the interview, the questions are prepared in advance and added to the research. The interview will be recorded, and a transcript of the interview is available for replication purposes. To ensure privacy of the experts, names are not entered in the transcript. Only general functions are used when needed.

Through the framework for the Design Science Research Method, a rigorous and relevant process is executed to contribute to existing research and problem statement.

4. Results

According to the DSRM model this section starts with Design and Development of the SCRUMBAN Data Science development model. Then the model will be demonstrated, and data will be extracted. The last step is to evaluate the data and present the results.

4.1. Design and development

The SCRUMBAN Data Science development model is based on the CRISP-DM development process combined with SCRUM and KANBAN elements. SCRUM and KANBAN give the best fit for Data Science projects (Saltz & Heckman, 2018)The base of the SCRUMBAN Data Science development model comes from SCRUM which provides a complete process with guidelines, rules, roles, events and artefacts, which makes it easier for development teams to adopt. They know exactly what is expected and what the process looks like. The elements of SCRUM that do not fit for Data Science projects are removed from the model. As a last step, elements that fit Data Science projects are added to the SCRUMBAN Data Science development model.

4.1.1. Roles:

The roles of the SCRUMBAN Data Science development model are based on SCRUM roles.

The development team consists of minimal three and a maximum of nine people, this is a rule from SCRUM (Schwaber & Sutherland, 2017). When the team consists of less than three members it lacks capacity to boost creativity and decreases interaction. Exceeding the maximum of nine people requires more coordination and communication which adds unwanted complexity. Every member of the team has his own specialization. One of the team members must be a Business Architect to control feasibility and sustainability of the proposed solution and the translation to business requirements. The architect is the only member that is not dedicated to one team only. The other members have only one team and must consist of different disciplines or with knowledge of different domains.

One member of the SCRUMBAN team is the Business Owner. He or she is the linking pin with the business. Responsibilities are: understanding business wishes and translating them into user stories and managing the backlog.

SCRUM master, he or she is of service to the team and to help the development team and product owner with their roles. They take away any obstructions for the development team on completing their tasks. All roles are taken from SCRUM because they give a clear understanding of the responsibilities of every member.

4.1.2. Events.

The events are based on SCRUM events. All events combined form the process of the development cycles. The goal of a SCRUM cycle is to deliver working software. The goal of cycle the SCRUMBAN Data Science development model is to improve on the process so that changes can be adapted quickly. Therefor not all events are suitable for the SCRUMBAN Data Science development model. To improve on the process, evaluation must be done on a regular basis. For the evaluation, trust and communication between team members is key. Only when there is trust, people will speak their minds. Another important aspect copied from SCRUM to the SCRUMBAN Data Science development model is the order and duration of the events. This sets a rhythm to the development; team members and stakeholders know what to expect and the rhythm brings stability to the process.

Sprint: A sprint loop is decided by the team in which they have several of the other events for the team only. Different to software development it is unclear in Data Science Projects if the goals can be achieved. Therefor working software is not the main goal of the sprint. Therefor a sprint in SCRUMBAN is to set a heartbeat for the team to have events planned. It only is to set a heartbeat for the team to have events planned. No events can be skipped, the reason is to get the most added value out of the SCRUMBAN Data Science development model. A sprint is set for three weeks.

Daily stand-up: every day the team takes between five and ten minutes to pitch what they achieved yesterday, what the planning is today and if they see any obstructions in the future. Here, it is possible for other team member to offer help and for the SCRUM master to act on the possible obstruction to clear them before they happen.

Sprint retrospective: At the end of every sprint, the team has a meeting within a secured environment. No-one else is allowed to attend this meeting. It's the most important meeting for everyone to be honest what went well and what needs adjustments. People should consider the rules of feedback to be respectful to each other. This meeting takes between 30 minutes and 1 hour. The outcome of this meeting is for the team to know how to improve themselves. It is not for anyone else than the team. The points to work on are points taken into the new sprint and are part of the next refinement.

Refinement: a set event at the beginning of the sprint. The product owner presents the new user stories on the backlog and discusses them on maturity. If the user story is understood and accepted by the team the product owner takes one sprint to prioritize the accepted user story. When an user story is not accepted, the product owner returns to the business to discuss the open issues from the development team. The aim is for the product owner to grow to get user stories clear before refinement starts.

Not all events of SCRUM are selected for the SCRUMBAN Data Science development model. The review is not selected for the SCRUMBAN Data Science development model. The goal for the review is to showcase working software to the business. Since the goal of the cycle of the SCRUMBAN Data Science development model is to improve on the process and the uncertainty of delivering working Data Science models at the end of the cycle, it doesn't seem to fit the purpose. Inviting the business and canceling every time the model isn't ready for review, contributes to the distrust of the business.

4.1.3. Artefacts

Artefacts are based on SCRUM, KANBAN and CRISP-DM. From SCRUM the Backlog, Increment and the User Story are copied into the SCRUMBAN Data Science development model. They give insight in the requirements and the priorities. From KANBAN the continuous workflow and the board to show Work-In-Progress are copied to the SCRUMBAN Data Science development model. These give insight for the team what they should work on, who works on what requirement, what the status of the Work-In-Progress is and how the work is balanced based on capacity. The CRISP-DM process provides the phases of the development. The phases combined with the board gives insight for the team and the stakeholder about the status of the work to deliver.

SCRUMBAN board: A board that is divided into different phases according to CRISP-DM and the user stories on the backlog. Within all the process steps all activities described in CRISP-DM are performed. Whenever a task is completed, the ticket moves to the next step. The ticket is filled with relevant information, so everyone is informed with one source only. In figure 5 the SCRUMBAN board is presented.

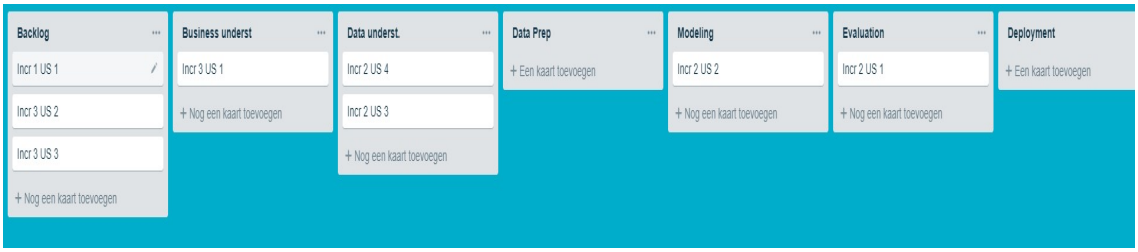


Figure 5: SCRUMBAN board based on the CRISP-DM development process.

Backlog: All user stories that are complete and prioritized. As long as the user stories are on the backlog priorities might change. When they move into the development process they are set.

Increment: Every user story belongs to an increment. An increment can consist of several user stories. An increment is a complete product for the business.

User story: The user story is weighted with points according to size and complexity. Every team has several points to work on, so they only work on user stories that can be managed by the team. For instance, the team can work on twelve points at once, so they can work on user stories that combine twelve points or less in the whole development process. The amount of points to work on is decided by the team. The points a team can divide, depends on knowledge, experience and size. The points can be adjusted every retrospective. During the sprint the team is not able to change the points assigned to them.

Important difference to SCRUM is the review. In CRISP-DM the review is part of the evaluation step of the process. In this step the solution is tested. When successful the solution is presented to the business. Here the business accepts or rejects the solution. When the solution is rejected the task returns to the backlog for refinement and prioritization. Time is only taken from the business when there is a solution to be reviewed. A review session must be attended by the stakeholders involved and can't be rejected by the business. Whenever a task is deployed or placed on the backlog the team has new points to spend. This improves continuous development.

This method is built around the twelve principles and the four values written in the Agile manifesto. In figure six the SCRUMBAN Data Science development modelis presented.

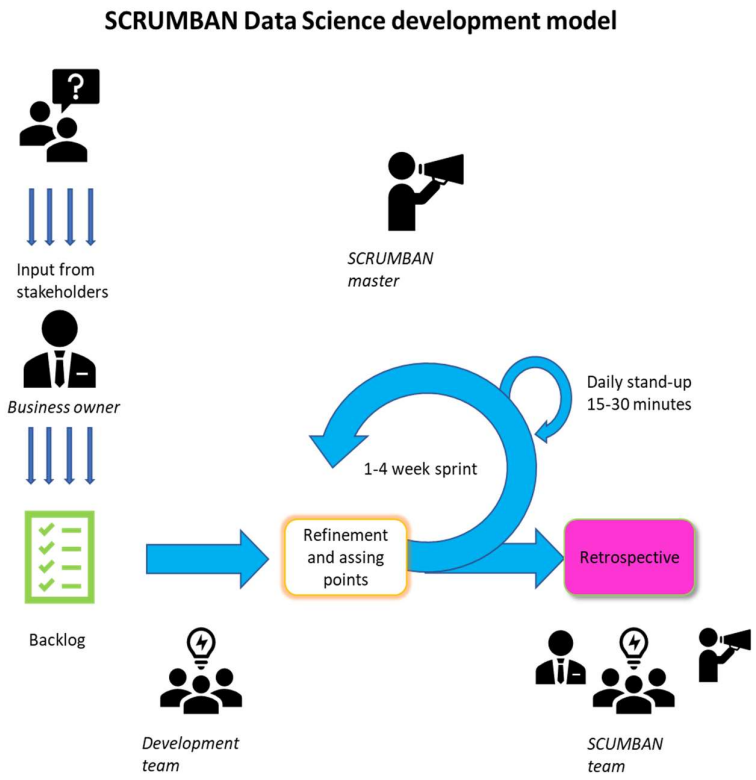


Figure 6: SCRUMBAN Data Science development model

4.2. Demonstration and data extraction

4.2.1. Demonstration

The selected company for the research of the SCRUMBAN Data Science development model is a large international consultancy company with a focus on global enterprises. The company is based in the Netherlands and employs 6,500 consultants. The company consists of different departments. The department where Data Science consultants are situated is Data and Insights. The department is divided in two sections. The first is focused on Business Intelligence and the other one is focused on Data Science, Artificial intelligence and Big Data. There are 140 consultants working for Data and Insights. The consultants have various years of experience, from ten years of experience in Data Science to the Young Professionals that just onboarded. The projects executed by the consultants come from existing clients. Due to the focus on Data Science projects, there is an increase in request from clients to deliver expert knowledge on these projects.

The demonstration of the SCRUMBAN Data Science development model is done in individual interviews with experts from different domains. Five consultants are selected for the interviews. There are three selection criteria defined for the selection of experts for the demonstration and interviews. The first is the years of experience, which must be over three years. The second selection criterion is the role they currently perform for a client. And third selection criterion is their experience with Agile development processes. The selection resulted in five consultants with four of them from the Data Science domain and one from Software Engineering with deep Agile knowledge. This role is added to the experts to give input on experience of effectiveness of Agile methods on different types of projects.

From the Data Science domain, the roles selected for the interviews are:

- One Data Scientist
- One Business Analyst
- One Techlead/Architect
- One SCRUM master

From the Software Engineering domain, the role selected is:

- Project manager

The purpose of the demonstration is to evaluate the SCRUMBAN Data Science development model on four criteria:

- Feasibility: the degree of practicality of the proposed model. Feasibility gives insight if the model could be used in a work environment. If the model is not feasible to begin with it has no purpose to exist.
- Completeness: the degree of having nothing missing on the model. Completeness shows if all important components are taken into account in order for the model to work.
- Usability: the degree to which the model is fit to be used. When the model is complete but not fit for use, the model will be rejected and has no reason to exist.
- Effectiveness: the degree to which the model is successful. In this context it is an estimation of the effectiveness of the model in theory. This criterion is connected to the research question.

The demonstration consists of a presentation of the SCRUMBAN Data Science development model by the researcher. For the presentation used in the demonstration see Appendix III. During the demonstration the consultants are guided by an interview. The interview consisted of questions that contribute to the evaluation criteria. The demonstration and interview take approximately one hour per consultant and are executed on individual basis. Due to scarcity of time from the consultants, three out of five interviews were executed through Skype. All interviews are recorded on a voice recorder and are stored as evidence. During the interviews the model was explained by the researcher. The consultant interviewed was free to ask questions or make remarks on the model. This resulted in an open discussion of the model.

After the interviews were executed the recordings were translated into transcripts. These transcripts contain unstructured data. Remarks are coded with timestamps to connect the scripts to the recordings. The unstructured data is then ready for extraction of data.

4.2.2. Data extraction

To answer the evaluation criteria accordingly, the unstructured data needs to be coded. Coding has been done by dividing the data into segments. The segments are connected to the evaluation criteria. The segments of the data are:

- Reflection on Roles
- Reflection on Events
- Reflection on Artifacts
- Reflection on the complete SCRUMBAN Data Science development model and SCRUMBAN board

To evaluate the SCRUMBAN Data Science development model the evaluation criteria are connected to the data segment that provides the information. Table 1 provides an overview which segments provide information per evaluation criteria.

Evalution Criteria	Segment
Feasibility	• Reflection on the complete SCRUMBAN Data Science development model and SCRUMBAN board
Completeness	• Reflection on Roles • Reflection on Events • Reflection on Artifacts • Reflection on the complete SCRUMBAN Data Science development model and SCRUMBAN board
Usability	• Reflection on Roles • Reflection on Events • Reflection on Artifacts • Reflection on the complete SCRUMBAN Data Science development model and SCRUMBAN board
Effectiveness	• Reflection on the complete SCRUMBAN Data Science development model and SCRUMBAN board

Table 1: Data segments connection to evaluation criteria

The remarks are then classified as either positive or negative remarks. This classification is needed to provide insights what is perceived to be successful parts of the SCRUMBAN Data Science development model and what parts need to be improved. For the extraction of the scripts see Appendix IV. The extraction and process of the data forms the basis of the evaluation of the SCRUMBAN Data Science development model.

4.3. Evaluation of the results

The SCRUMBAN Data Science development model evaluation is divided into four criteria. Per criteria the positive and negative remarks from the data extraction are considered. In this section the results are presented per evaluation criteria.

Feasibility: the degree of practicality of the proposed model.

Feasibility is the evaluation of the practicality of the proposed SCRUMBAN Data Science development model. All consultants reply that the SCRUMBAN Data Science development model could work though testing of the SCRUMBAN Data Science development model is needed. The SCRUMBAN Data Science development model provides an out of the box process with roles, events and artifacts just like SCRUM. The SCRUMBAN Data Science development model is tweaked for the purpose of Data Science. A positive remark of one of the consultants is: “Good model, use it and try it out within a company, it’s a model to test”. Another remark about the feasibility of the model is: “It’s a model that could work, an experiment gives insights”. These two remarks do not directly give answer on feasibility, but they give insights if the model is easy to understand. How easy a model is to understand is directly connected to feasibility.

Completeness: the degree of having nothing missing on the model

Completeness gives insights if all the elements that should be in the model, are provided. To give answer on completeness the SCRUMBAN Data Science development model is evaluated on all elements separately and on the complete SCRUMBAN Data Science development model. On roles most discussion was on the role of the architects placed within the SCRUMBAN team. “Having an

Architect close to the development team could result in the solution being stopped to early before feasibility is tested", according to one of the consultants. Another remarked "Having an Architect close to the development team could result in the solution being reviewed quickly". Though the role of the Architect for the SCRUMBAN team must be defined properly and according to several consultants it is not necessary to be part of the team. A consulting role for the Architect should suffice. "Architect a good idea but questions about feasibility to add one architect per team. At client solved by one dedicated point of contact from a pool of Architects."
There is consensus between all consultants about the size of the team. "Not more than nine people is important to keep the learning capability of the team and a secure environment. When the team gets bigger the secure environment is lost". Though the roles within the teams are not very clear for all consultants. They all think the team should have at least three Data Scientists. The Architect mentioned adding a tester to the team to make it more complete.
Next topic was the events considered. All consultants replied that the events are important for the team to grow and become highly effective. Therefore, they all concluded that the retrospective is the most important event in the cycle and can't be skipped. Though it becomes clear that an experienced SCRUMBAN master is needed. "Retrospective needs an experienced SCRUMBAN master. Uses different techniques to get information from the team. Success is dependent on the experience of the SCRUMBAN master. The team members should feel at ease in order to tell what they think". Trust between team members form a very important base for all activities. The experience is that some people can't give feedback, especially when there is no trust between team members. It's the task of the SCRUMBAN master to help the team to create this trust.
The artifacts are not clear for everyone. Especially the points to divide took time to explain. The Project Manager asked about a wall of reference to help the team to assign points to the new user stories. "Do you use a wall of reference to get an indication on US point? Referencing to previous work. The team knows what the value of a point is and how an US is weighted". It is for the team only and has no other purpose than to help the team to weigh User Stories.
Overall the remarks on the SCRUMBAN Data Science development model were positive. The Architect mentioned to assign a goal to each sprint, so the team knows what they are going to achieve, even though it is hard for Data Science projects to create goals when the target is uncertain. Goals for a sprint could be to move a certain User story into another phase or to work on improvements coming from the retrospective. All consultants mentioned the SCRUMBAN Data Science development model should be tried and tested to see if it works in a normal production environment.

Usability: the degree to which the model is fit to be used

Overall the consultants found the SCRUMBAN Data Science development model easy to understand. The roles are complete and clear. They are copied from SCRUM and well known by the consultants. How the development team is combined depends on the project and its requirements. The events and the purpose of the events are clear and understood by the consultants. No questions were raised by the consultants. On the artefacts that were presented some minor questions arise about the new SCRUMBAN Data Science development model. The assignment of points needed more explanation for the concept to land. Some consultants gave advice to make this easier and more understandable by adding an artifact to the SCRUMBAN Data Science development model or to divide the SCRUMBAN board in another way. None of the consultants replied that the SCRUMBAN Data Science development model was too complex to understand. All consultants replied that the SCRUMBAN Data Science development model should be tested to see if it is usable. From a

theoretical perspective the SCRUMBAN Data Science development model is gives a complete process with roles and rules.

Effectiveness the degree to which the model is successful

Some of the consultants have deep Agile experience since more projects are executed through Agile processes. They are part of the Data Science division of the company and more clients are experimenting with Agile processes. To stay current, most consultants followed a SCRUM course or are certified SCRUM masters. Although they have Agile experience, they approach the SCRUMBAN Data Science development model from their Data Science project experience. They know what works and what doesn't work. Most of the consultants expect the SCRUMBAN Data Science development model to be successful when it is used. It is emphasized that trust between team members is needed to become a high-performance team. This can't be achieved in one day and takes time. Stability of the team, they should work together for a long period of time. Create good environments to secure successful teams. How a team is composed and how it works together gives the difference between a high-performance team and an average team. The success factor of a team is dependent on how a team works together. They should believe in the SCRUMBAN Data Science development model and use and tweak the model, so it works. "Adapt the complete model. The basis (Cycle) should be completely executed. Changing the model will result in failure, spoken from experience.". There is also a bottleneck highlighted by the Business Analyst: "A bottleneck could arise in modelling phase. Business owners could get nervous which stands in the way of implementation of the model. Think about decreasing points when User Stories stay longer in a phase. Or use the strength of Agile to help each other".

The results of the evaluation of the SCRUMBAN Data Science development model are listed in table 2

	Roles	**Events**	**Artefacts**	**Model**
Feasibility	Not reviewed	Not reviewed	Not reviewed	The model is easy to understand which makes it feasible and suitable for purpose
Completeness	There is no need for an Architect being part of the team. An architect available for consult is enough.	Complete, no events were missing.	Point system is too complex, use a wall of reference. Assign a sprint goal for the team to reach. This gives focus.	Complete, there were no suggestions for additions to the model other than mentioned in roles and artefacts
Usability	Roles and responsibilities are clear and usable.	Events and goals of the events are clear and usable.	Point system too complex and therefore, not usable. Replace it by a wall of reference which easy to understand.	Complete, the model is easy to understand and complete with a process, roles and rules.
Effectiveness	Not reviewed	Not reviewed	Not reviewed	The model is perceived to be effective if all team

				members and stakeholders believe in it. And it is important that the complete process, roles, events and artefacts are deployed, skipping one part had a negative effect on the performance of the SCRUMBAN Data Science development model.

Table 2: Overview of the results per criterion

The results of the review by the constants give a clear overview of improvements on the SCRUMBAN Data Science development model. This results in the following model:

Roles:

A development team with a minimum of three members and a maximum of nine members to support creativity and avoid complexity. The development team consists of Data Scientists, Business Analysts, User Experience Expert and Data Engineers. The combinations of roles depend on the required solution.

A Business owner to represent the business, clarify requirements and prioritize work to be done.

A SCRUMBAN master as a facilitator to the team and stakeholders.

Next to this dedicated team there should be an Architect available on consulting basis, to review the required and proposed solution against the development principles of the company.

Events:

Sprint: a process chain that consist of several events. Every sprint should start with a goal for the team to reach by the end of the sprint. This can be goals on work to deliver or process improvements that are executed and evaluated. This gives the team focus. The duration of a sprint is between one and four weeks and is set by the team.

Daily stand-up: to gain insight what every individual is working on and to tackle impediments quickly. The stand-up should take between fifteen and thirty minutes.

Sprint retrospective: the team looks back on the sprint and give each other feedback. The team decides what parts of the process went well and what to improve. The points to improve are taken into the next sprint. The goal is to improve on the process and to gain trust between team members. The retrospective should take between one and two hours.

Refinement: at the beginning of the sprint the team sits together to discuss new user stories and to understand the priorities of the user stories. This is also the moment the team sizes the workload of the different user stories according to the wall of reference.

Artefacts:

SCRUMBAN board: a board to gain insight of the backlog and work-in-progress. The CRISP-DM phases are plotted on the board and work is pulled into the phases. There is a continuous workflow of user stories in progress. The actual status of work is available for the team and stakeholders.

Backlog: all user stories that are complete and prioritized. As long as the user stories are on the backlog priorities might change. When they move into the development process they are set.

Increment: every user story belongs to an increment. An increment can consist of several user stories. An increment is a complete product for the business.

User story: a user story gives context and purpose to the requirement. The user story is sized against previous user stories. The size in combination with the available team member and the phase, gives assurance that there is enough work-in-progress to handle without the risk of an overload of work-in-progress.

Wall of reference: a reference guide for the team based on previous user stories to size new user stories. This is based on experience and simple to use. How the size is measured, is decided by the team. Example of size are T-shirt sizes S, M, L, XL or hours. The wall of reference can be adjusted over time when the team gains speed.

Figure 7 shows the adjusted SCRUMBAN Data Science development model

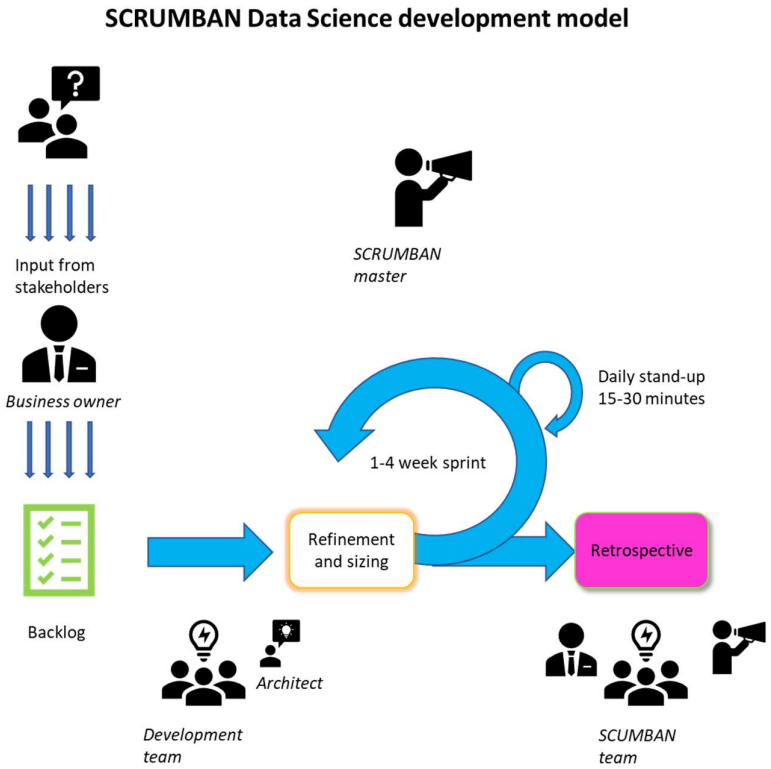


Figure 7: Adjusted SCRUMBAN Data Science development model

The results give answer to sub question three: **Is the proposed model suitable for Data Science projects?** The answer is yes. The proposed changes have minor impact on the original SCRUMBAN Data Science development model. The change was on a role taken out of the team and an easier

way to weigh the work to be done. The model fits Data Science projects since the process is based on the phases of CRISP-DM, which is designed for Data Science projects. All the consultants replied that the SCRUMBAN Data Science development model should be tested in a real Data Science project, to see if it works.

5. Discussion, conclusions and recommendations

This chapter gives answer to the research sub question. To answer this question the results of the design research are discussed, and a conclusion is drawn. And to complete the chapter recommendations of further research are given.

5.1. Discussion

The SCRUMBAN Data Science development model is perceived to be feasible. There is no substantial remark made during the interview that implies that the SCRUMBAN Data Science development model wouldn't sustain in a real team. The SCRUMBAN Data Science development model is easy to understand, and it is a complete process with roles, events and artefacts. It is based on SCRUM which is well known in the consulting world. The literature already proved Agile to be effective for system engineering and the remarks of the project manager helped to give an experienced view on the SCRUMBAN Data Science development model. It is easy to understand and ready for use. Therefore, it is concluded that the SCRUMBAN Data Science development model is feasible.
Almost all elements of the SCRUMBAN Data Science development model were clear for the consultants. When looking at the roles that are described most consultants agreed on at least three Data Scientists, a Data Engineer and a User Experience developer. The three Data Scientist are needed to give each other feedback on work or help with difficult tasks or requirements. Though the project manager isn't aware of the roles that are needed for Data Science projects, it was emphasized not to exceed the limit of nine team members. This is to ensure the trust between team members and to become a high-performance team. It is needed to have an Architect near the team to give guidance on directions and to test if the solution meets the company goals and rules. But the Architect doesn't need to be part of the team. It is likely to be feasible for every team to have an Architect since this is not a well-known role within the company. If there is such a role it is likely to have a mismatch in resources and work to be done. This is inefficient from the Architects perspective as well as the team perspective.
On events there was one remark if the review session could be added to the sprint cycle to get feedback on work that has been done, even though the development process has a testing and review phase as well. This could be a peer review instead of a business review. Other consultants thought the stand-up gave enough space for other team members to give feedback or to ask for help. The review as a part of the process is sufficient.
Important outcome of the demonstration and interviews was the unclarity about the point system to prevent too much work for a team to handle. The project manager, therefore, suggested to add a wall of reference to the SCRUMBAN Data Science development model as an artefact. The wall of Reference is for the team only and gives insight how to weigh a certain User Story based on experience. This is an easier method then a point system.
The SCRUMBAN board is a nice add to the SCRUMBAN Data Science development model to give insight for the team and stakeholders on work in progress. Since it is divided into CRISP-DM phases the SCRUMBAN Data Science development model is set for Data Science development purposes. Based on these outcomes it is concluded that with some small adjustments the SCRUMBAN Data Science development model is complete.
Furthermore, the SCRUMBAN Data Science development model is perceived to be usable and practical model, and it is based on known Agile models. In literature these models are proven to be effective. The Project Manager agrees that Agile models and especially SCRUM works, but remarks that all elements of the model should be executed. Missing or skipping one or more elements will have a direct effect on usability and effectiveness of the model. This is also agreed by the Architect

and Business Analyst. The model is very clear and when adjustments have been made on roles it could be tested in practice.

Though it is easy to understand there is also a risk in the SCRUMBAN Data Science development model. This is a model that needs to mature to gain trust. Agile works due to trust between team members. This trust is not there when the team starts, but it must grow. With every event the experience and trust of the team members grows. As remarked by one of the consultants is that it's not needed to have high potential team member to be a high-performance team. Trust makes this difference. Therefore, it takes time to measure if the SCRUMBAN Data Science development model is effective. Based on Agile experience of the consultants the SCRUMBAN Data Science development model is perceived to be effective, though it should be tested for at least two to three months before a conclusion can be drawn.

This research also contributes to existing literature. The articles that were found, gave insight about previous attempts for combining Agile methods with Data Science development processes. This research continued by creating a model from successful elements of Agile methods that were suitable for Data Science projects. Elements that didn't add value to the process were skipped. SCRUMBAN is known by consultants but hasn't been researched yet on effectiveness for Data Science projects.

5.2. Conclusions

In conclusion, some of the elements of the SCRUMBAN Data Science development model need to be adjusted for the model to be complete. This could have influence on the performance and effectiveness of the team. There are some important remarks that all components of the SCRUMBAN Data Science development model should be executed for the model to work. Skipping an event or creating a bigger team will result in lesser growth in trust and therefore, missing out on achieving improvement on performance.

Based on the results an answer can be given to the research question: **How to incorporate the Agile methodology into Data Science projects to gain flexibility?** First the theoretical framework was built from existing research on Agile and Data Science. First the development processes for Data Science projects were reviewed. CRISP-DM had an advantage over KDD, and therefore, it was decided CRISP-DM was more suitable to fit with Agile methods. Secondly the research focussed on Agile. It became clear which values and principals form the basis of Agile, and the models that fit Data Science projects. Attempts of fitting Agile methods were made with SCRUM and KANBAN. SCRUM provides a complete process with roles, events and artefacts. KANBAN creates a continuous workflow of Work-in-progress. From these two Agile methods, successful elements were selected and combined with the CRISP-DM process which resulted in the SCRUMBAN Data Science Development model. This model was reviewed on four evaluation criteria by experienced Data Science consultants. These criteria were: feasibility, completeness, usability and effectiveness. As a result, the model was adjusted with two recommendations. The first recommendation was to keep the Architect separate from the development team. And the second recommendation was to simplify the method for weighing the work to be done. All consultants replied that they thought the SCRUMBAN Data Science development model is fit for purpose. Therefor the conclusion is that the SCRUMBAN Data Science development model contains is a complete process, which is easy to understand and ready for use. The model gives insight on work-in-progress, that makes it easy to point out problems and changes in an early stage. The team can therefor act quickly, which results in more flexibility.

5.3. Recommendations for practice

The SCRUMBAN Data Science development process is concluded to be feasible and usable. It hasn't been researched on effectiveness, but remarks were made that the process is complete and ready for use. Agile and especially the SCRUM method is proven to be effective. The elements from KANBAN gave insight in status quickly. Being based on these two methods and only keeping the elements that add value make the SCRUMBAN Data Science development model suitable to try out. Every element has a clear rule and goal. The model is easy to understand and can easily be deployed. It is recommended to try the SCRUMBAN Data Science development model in a small setting to check if the expectations are met. When the SCRUMBAN Data Science development model has proven to be effective and fir for purpose, it can be deployed to a larger scale.

5.4. Recommendations for further research

There are two ways to research the SCRUMBAN Data Science development model further. First is get a review on the SCRUMBAN Data Science development model from another perspective, namely the business. The model is reviewed by consultants only and not in a business setting. Review from a business perspective could give other results. In consultancy, Agile methods are used heavily and well know. In companies, and especially Data Science departments, Agile methods are less known and therefore, more explanation could be needed. Also, responses could differ from those of consultants. Secondly the SCRUMBAN Data Science development model can be tested with real Data Science Projects. As stated by all the consultants, it is a model to be tested. This could be done as a case study within a company that has experience with Data Science projects. The primary research question still needs to be answered. Flexibility creates the ability to adapt quickly to changes. Flexibility of the SCRUMBAN Data Science development model can only be tested by using the model in practice. The SCRUMBAN Data Science development model should at least be tested for more than five cycles and set against development duration in a waterfall setting. This is the baseline against which the speed of every development is compared. The results of every measured development duration needs to be evaluated, so a conclusion can be drawn. When the development time speeds up, it could result in more flexibility.

6. Reflection on the research

First part of the research took some iteration to get to a good research question. When the research question was formulated correctly, it became easier to find relevant research papers to build a theoretical framework and create a model.
The guidance from the university was sufficient to give you tips on the way to proceed but also gave you enough opportunity to have your own thoughts how the research could be conducted.
A challenge arose when I lost of my research company. Without notice they pulled out of the research. Fortunately, the consulting company stepped in. Conducting the interviews had the same time issues as mentioned before. And it was a struggle since the consultants selected for the interviews were not all located at the same office. Most of them were at clients so the interviews were conducted through Skype. Next time I would not conduct interviews through Skype since you miss facial expression since the presentation took over from the camera. Therefore, I could not see the interviewed consultant and missed the facial expression which could give extra information. The preparation of the interviews helped to check if all evaluation criteria were touched. And I noticed that when the interviews are done separately, results of experiences from the previous interviews changed the next one a bit. Next time I would consider getting the group together and conduct the interview at the same time, but it would introduce the problem of group bias. This could also help the respondents to think about other responses and give other insights. Now it is mainly one-sided. Furthermore, I noticed that in some interviews it was needed to take the lead while other interviews the respondents gave a lot of information themselves.
This was the first time I used the Design Science Research Method. I thought it was a good way to reflect on a practical question. The model also helped the consultants with a tangible model instead of a hypothetical model. It gave more body to the review and it took less time to explain the model. It is easier to reflect on a proposition than to reflect on a model that only exists in theory. I was happy with the results from the review and it added value to the SCRUMBAN Data Science development model. The improvements made it easier to implement.
Overall, I think the research went well and I could tackle the loss of the research company on time.

References

A list of all references used, in accordance with the APA format.

Asay, M. (2017, November). 85% of big data projects fail, but your developers can help yours succeed. *TechRepublic*. Retrieved from https://www.techrepublic.com/article/85-of-big-data-projects-fail-but-your-developers-can-help-yours-succeed/

Avison, D. E., & Fitzgerald, G. (2003). Where now for development methodologies? *Communications of the ACM*. https://doi.org/10.1145/602421.602423

Beck, K., Beedle, M., Bennekum, A., Cockburn, A., Cunningham, W., Fowler, M., … Thomas, D. (2001). Manifesto for Agile Software Development. https://doi.org/10.1177/0149206308326772

Cao, L. (2015). Data Science: A Comprehensive Overview. *ACM Computing Surveys*, *50*(3).

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0 (Step-by-step data mining guide)*. *CRISP-DM Consortium*. https://doi.org/10.1056/NEJMoa1108524

Dhar, V. (2013). Data Science and Prediction. *Communications of the ACM*. https://doi.org/10.2139/ssrn.2086734

do Nascimento, G. S., & de Oliveira, A. A. (2012). An Agile Knowledge Discovery in Databases Software Process. *LNCS*, 56–64. https://doi.org/10.1007/978-3-642-34679-8_6

Dybå, T., & Dingsøyr, T. (2008). Empirical studies of agile software development: A systematic review. *Information and Software Technology*. https://doi.org/10.1016/j.infsof.2008.01.006

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*. https://doi.org/10.1145/240455.240464

Grigoriev, P. A., & Yevtushenko, S. A. (2003). Elements of an Agile Discovery Environment. In *Discovery Science*. https://doi.org/10.1007/b14292

Kisielnicki, J., & Misiak, A. M. (2016). Effectiveness of agile implementation methods in business intelligence projects from an end-user perspective. *Informing Science*. https://doi.org/10.1515/fman-2017-0021

Kisielnicki, J., & Misiak, A. M. (2017). EFFECTIVENESS of AGILE COMPARED to WATERFALL IMPLEMENTATION METHODS in IT PROJECTS: ANALYSIS BASED on BUSINESS INTELLIGENCE PROJECTS. *Foundations of Management*. https://doi.org/10.1515/fman-2017-0021

Larson, D., & Chang, V. (2016). A review and future direction of agile, business intelligence, analytics and data science. *International Journal of Information Management*. https://doi.org/10.1016/j.ijinfomgt.2016.04.013

Lei, H., Ganjeizadeh, F., Jayachandran, P. K., & Ozcan, P. (2017). A statistical analysis of the effects of Scrum and Kanban on software development projects. *Robotics and Computer-Integrated Manufacturing*. https://doi.org/10.1016/j.rcim.2015.12.001

Marbán, O., Segovia, J., Menasalvas, E., & Fernández-Baizán, C. (2009). Toward data mining engineering: A software engineering approach. *Information Systems*. https://doi.org/10.1016/j.is.2008.04.003

Mariscal, G., Marbán, Ó., & Fernández, C. (2010). A survey of data mining and knowledge discovery

process models and methodologies. *The Knowledge Engineering Review*. https://doi.org/10.1017/S0269888910000032

Muntean, M., & Surcel, T. (2013). Agile BI: The Future of BI. *Informatica Economica*. https://doi.org/10.12948/issn14531305/17.3.2013.10

Naur, P. (1966, July). The science of datalogy. *Communications of the ACM*, 485.

Okoli, C., & Schabram, K. (2010). A Guide to Conducting a Systematic Literature Review of Information Systems Research. *Working Papers on Information Systems*. https://doi.org/10.2139/ssrn.1954824

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*. https://doi.org/10.2753/MIS0742-1222240302

Royce, W. W. (1987). Managing the development of large software systems: concepts and techniques. *Proceedings of the 9th International Conference on Software Engineering*, 328–338.

Saltz, J. S., & Heckman, R. R. (2018). A Scalable Methodology to Guide Student Teams Executing Computing Projects. *ACM Trans. Comput. Educ.*, *18*.

Saunders, M., Lewis, P., & Thornhill, A. (2016). *Research Methods for Business Students*. *Research methods for business students*. https://doi.org/10.1111/j.1365-2222.2005.02180.x

Schmidt, C., & Sun, W. N. (2018). Synthesizing Agile and Knowledge Discovery: Case Study Results. *Journal of Computer Information Systems*. https://doi.org/10.1080/08874417.2016.1218308

Schwaber, K., & Sutherland, J. (2017). The Scrum Guide. https://doi.org/10.1053/j.jrn.2009.08.012

Serrador, P., & Pinto, J. K. (2015). Does Agile work? - A quantitative analysis of agile project success. *International Journal of Project Management*. https://doi.org/10.1016/j.ijproman.2015.01.006

Walker, J. (2017, November). Big data strategies disappoint with 85 percent failure rate. *Digital Journal*. Retrieved from http://www.digitaljournal.com/tech-and-science/technology/big-data-strategies-disappoint-with-85-percent-failure-rate/article/508325

Wells, D. (2013). A Gentle Introduction. *Extreme Programming*. https://doi.org/10.1111/1467-9973.00225

Appendix I Selected articles

The table below gives an overview of the selected articles for data extraction. This data is used to build the theoretical framework. The overview is sorted by year of publication. The sorting is done to give insights how research on Agile and Data Science has evolved.

Title	**Author(s)**	**Year of publication**
Elements of an Agile Discovery Environment	Grigoriev, Peter A Yevtushenko, Serhiy A	2003
Agile BI – The Future of BI	Muntean, Mihaela Surcel, Traian	2013
A review and future direction of agile, business intelligence, analytics	Larson, Deanne Chang, Victor	2016
Effectiveness of Agile Implementation Methods	Kisielnicki, Jerzy Misiak, Anna Maria	2016
Effectiveness of Agile compared to waterfall implementation methods in IT projects	Kisielnicki, Jerzy Misiak, Anna Maria	2017
Future software organizations – agile goals and roles	Kettunen, Petri Laanti, Maarit	2017
A Scalable Methodology to Guide Student Teams Executing Computing projects	Saltz, Jeffrey S. Heckman, Robert R.	2018

Appendix II Interview questions

Below the question can be found that supported the interview. Another purpose for the interviews is to make sure all topics for evaluation of the model were addressed.

Before presentation of the model:

1. Introduce yourself?
2. What is your experience with Agile? (knowledge)

After explanation of the model:

3. Are any roles, events or artefacts missing?
4. What are the important elements from the model?
5. What is your experience with Agile, what works and what doesn't work?
6. What is your reflection on the presented model?

Appendix III Presentation of the SCRUMBAN Data Science development model

Data Science goes Agile

THE SCRUMBAN MODEL

MARISKA BLASWEILER

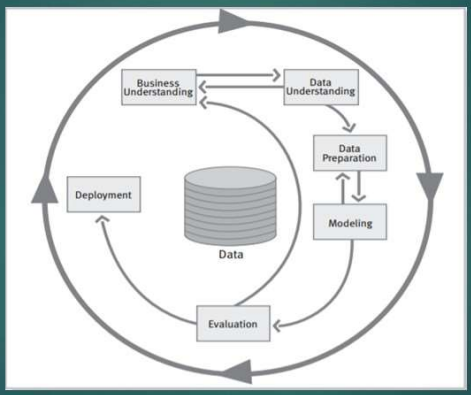
Problem definition

- Requirements and goals are not always set correctly for Data Science developments and there is a need for a development process that gains insights on a frequent basis and quickly adapts changes. done

Research question

- **How to incorporate the Agile methodology into Data Science developments to gain flexibility?**
  - SQ1. What are Agile values and methods and which one(s) will fit Data Science development processes? This question will be answered by literature study.
  - SQ2. Which Data Science development models are available, which attempts have been made to incorporate Agile into the development models? This question will be answered by literature study and will result in a proposed Agile Data Science development model.
  - SQ3. Is the proposed model suitable for Data Science projects? This question will be answered by a focus group interview.

CRISP-DM



Agile: four values

- Individuals and interactions over processes and tools
- Working software over comprehensive documentation
- Customer collaboration over contract negotiation
- Responding to change over following a plan

Agile: Twelve principles

- Satisfy the customer through early and continues delivery of valuable software
- Welcome changing requirements, even late in development. Agile processes harness change for the customer's competitive advantage.
- Deliver working software frequently, from a couple of weeks to a couple of months, with a preference to the shorter timescale.
- Business people and developers must work together daily throughout the project.
- Build projects around motivated individuals. Give them the environment and support they need, and trust them to get the job done.
- The most efficient and effective method of conveying information to and within a development team is face-to-face conversation.

Agile: Twelve principles

- Working software is the primary measure of progress.
- Agile processes promote sustainable development. The sponsors, developers, and users should be able to maintain a constant pace indefinitely.
- Continuous attention to technical excellence and good design enhances agility.
- Simplicity--the art of maximizing the amount of work not done--is essential.
- The best architectures, requirements, and designs emerge from self-organizing teams.
- At regular intervals, the team reflects on how to become more effective, then tunes and adjusts its behaviour accordingly.

SCRUMBAN the model *Roles*

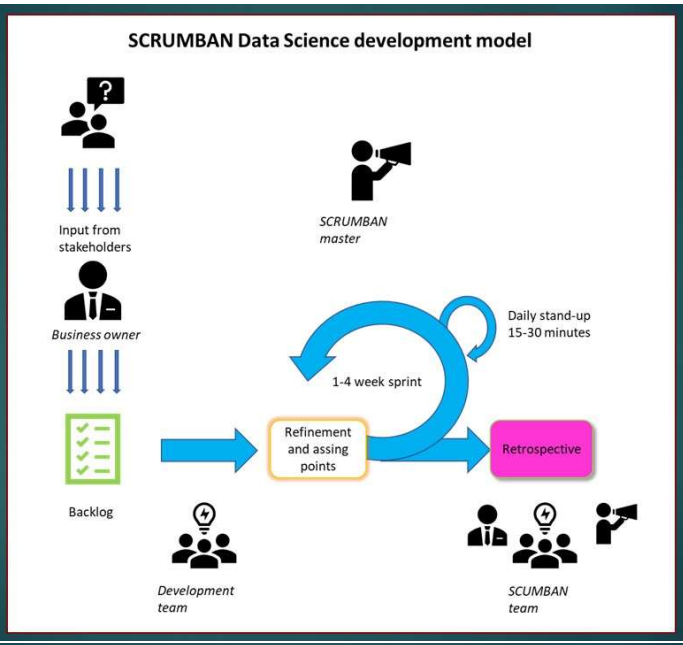
- The development team consists of minimal 3 and a maximum of 9 people. One of the team members must be an Architect to control feasibility and sustainability of the proposed solution.
- One member of the scrumban team is the Business Owner.
- SCRUM master

SCRUMBAN the model *Events*

- Sprint
- Daily scrum
- Sprint retrospective
- Refinement

SCRUMBAN the model *Artefacts*

- SCRUMBAN board
- Backlog
- Increment
- User story



SCRUMBAN Board

https://trello.com/b/lsqn6X7z/scrumban-crisp-dm

Appendix IV Data extraction table

In this table the remarks made during the interviews are collected and sorted on the reflection on the roles, events, artefacts and the model. The marks are divided into positive remarks that support the model and marks that are negative or suggestions on improvements. Some of the remarks are placed between quotes which means they are actual quotes made by the consultants. The other remarks are collections, summaries or interpretations of remarks.

Segment	**Positive remarks**	**Improvements or negative remarks**
Reflection on Roles		
	Architect is good to add, now missing (Data Scientist)	Role of architect needs more explanation; text is not very clear what is expected of the architect.
	"Kanban gives more freedom when the outcome is unclear" (Data Scientist)	There should be at least 3 data scientists in the team to avoid tunnel vision. (Data Scientist)
	"Having an Architect close to the development team could result in the solution being reviewed quickly". (Business Analyst 06:40)	A designer/ UX consultant and a data engineer is needed in the team. (Data Scientist)
	SCRUMBAN master should be independent from IT. A real facilitator". (Techlead/Architect 12:00)	"Having an Architect close to the development team could result in the solution being stopped to early before feasibility is tested". (Business Analyst 06:40)
	There should be someone who understands the domain. A business analyst in the team. One Data Engineer, one UX and at least three Data Scientists. (Techlead/Architect 15:00)	"The translation between business requirements and resources should be the responsibility of the architect". (Techlead/Architect 10:00)
	"Not more than nine people, is important to keep the learning availability of the team and a secure environment. When the team gets bigger the secure environment is lost". (Project Manager 10:00)	Maybe a tester is missing in the team. (Techlead/Architect 17:00)
	Experience that most of the times a dedicated SCRUMBAN master is needed. It's the best way to work. (Project Manager 12:30)	"Every team lacks a person that helps with platforms and needed tooling. So, all skills are available for the teams to achieve their goals." (SCRUM master 09:52)
	"Architect a good idea but questions about feasibility to add one architect per team. At client solved by one dedicated point of contact from a pool of	

	Architects." (SCRUM master 08:45)	
Reflection on Events		
	"Retrospective is the most important event. It is the strength and core of Agile. Most projects are still waterfall, but the experience is that Agile gives more capabilities to adapt. Also, stand-ups are important, it shows quicker when impediments arise". (Project Manager 15:00)	Don't forget the sprintplanning. Planning on targets is hard because it's unclear. (Techlead/Architect 23:30)
	"Retrospectives are important to give feedback on the process. Not everyone is able to give feedback". (Techlead/Architect 20:00)	"Retrospective needs an experienced SCRUM master. Uses different techniques to get information from the team. Success is dependent on the experience of the SCRUM master. The team members should feel at ease in order to tell what they think". (Business Analyst 14:55)
		"the sprint planning is missing from the events. Maybe needed, depending on the board." (SCRUM master 14:35)
Reflection on Artefacts		
		More explanation on points or more the Kanban way and work per person not per team. (Data Scientist)
		"Do you use a wall of reference to get an indication on US point. Referencing to previous work. The team knows what the value of a point is and how an US is weighted". (Project Manager 19:30)
		There could be dependency between US. You should make this clear in the refinement session. (Project Manager 27:00)
		Show when a US story or a phase is done (SCRUM master)
		Add a product backlog, this gives a team insight what they work on. (SCRUM master)
Reflection on complete model and SCRUMBAN board		

	"Easy to identify bottlenecks". (Business Analyst 24:00)	There is a dare. Most Data Scientist work within their own bubble. Most companies don't work on specific Data Science projects. Working in a structured way is not in place at most companies. Data preparation takes a lot of resources because data is not prepped properly. The structure of the model offers a structured way of working. (Business Analyst 29:06)
	"It's a model that could work, an experiment gives insights". (Data Scientist)	"A bottleneck could arise in modelling phase. Business owners could get nervous which stands in the way of implementation of the model. Think about decreasing points when User Stories stay longer in a phase. Or use the strength of Agile to help eachother". (Business Analyst 31:05)
	"This is a model that should be tried for at least two-three months. And should just be tried". (Business Analyst 38:00)	"A peer review moment should be added to the sprint with the remark that works doesn't have to be finished". (Data Scientist)
	"From an Architect point of view, you get grip on the matter at hand. Impediments are quickly shown". (Techlead/Architect 36:00)	How a team is composed and how it works together gives the difference between a high-performance team and an average team. The success factor of a team is dependent on how a team works together. They should believe in the model and use and tweak the model, so it works. (Business Analyst 34:00)
	"Good model use it and try it out within a company, it's a model to test". (Techlead/Architect 40:00)	Assign names or pictures of the people working on the User Story. (Techlead/Architect 37:00)
	The model should work for the situation as described for Data Science projects. (Project Manager 36:00)	Add a sprint goal to the sprint: A goal gives extra motivation to the team. With points coming from the retrospective. (Techlead/Architect 39:00)
		Stability of the team, they should work together for a long period of time. Create good environments to secure

		successful teams. (Project Manager 32:00)
		Adapt the complete model. The basis (Cycle) should be completely executed. Changing the model will result in failure, spoken from experience. (Project Manager 33:00)
	The model fits the purpose, Data Science projects are harder to plan (SCRUM master)	
	"The rhythm of SCRUMBAN is very pleasant." (SCRUM master 30:15)	